Elsevier LaTeX template

Elsevier[1]

Radarweg 29, Amsterdam

Elsevier Inc[1,1], Global Customer Service[1,]

[a] *1600 John F Kennedy Boulevard, Philadelphia*
[b] *360 Park Avenue South, New York*

Abstract

This template helps you to create a properly formatted LaTeX manuscript.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template

2010 MSC: 00-01, 99-00

1. Introduction

Since the dawn of internet and world wide web, humanity has witnessed a degree of connection beyond reckoning. The proliferation of digital devices pervaded with various applications that account for almost all aspect of humanity, have created cyber communities that constantly mutate [**?**]; [**?**]. In a world where we have network infrastructures that can support up to 250Mbps of data transmission, and smart phones and IOT devices that can have processing power of up to 3 Ghz, data becomes ubiquitous, the quantum that lays the foundation of the nexus [**?**].

According to InternetLiveStates.com [**?**], only in one second, there are 9,878 tweets sent, 1,138 instagram photos uploaded, 3,117,720 emails sent, 99,738 Google searches made, and 94,144 Youtube videos viewed. That is, if it has taken 5 second the read the preceding paragraph, during that time, 15,588,600 emails are sent.

*Fully documented templates are available in the elsarticle package on CTAN.

*Corresponding author

[1] Since 1880.

Driven by the ambition to harness the power of this deluge of data, the term 'Big Data' (BD) was coined [**?**]. BD initially emerged to address the challenges associated with various characteristics of data such as velocity, variety, volume and variability [**?**]. BD is the practice of extracting patterns, theories, and predictions from a large set of structured, semi-structured, and unstructured data for the purposes of business competitive advantage [**?**]; [**?**]. BD is a game-changing innovation, heralding the dawn of a new data-oriented industry.

Nonetheless, BD is not a magical wand that can enchant any business process. While a lot of opportunities exist in BD, subsuming an emergent and rather high-impacting technology like BD to current state of affairs in organizations, is a daunting task. According to recent survey from Databricks, only 13% of the organizations excel at delivering on their data strategy [**?**]. Another survey by NewVantage Partners indicated that only 24% organization have successfully gone data-driven [**?**]. This survey also states that only 30% of organizations have a well established strategy for their big data endeavour. In addition, surveys from McKinsey & Company ([**?**]) and Gartner ([**?**]) further support these numbers, which illuminates on the scarcity of successful big data implementations in the industry.

Among the challenges of data adoption perhaps the most highlighted are 'data engineering complexities', 'big data architecture', 'rapid technology change', 'lack of sufficient skilled data engineers', and 'organization's cultural challenges of becoming data-driven' [**?**];[**?**]. This focus of this study is on data engineering complexities and in specific big data architecture.

In the past, organization relied on a few technology giants to provide infrastructure and tools necessary for big data, while today there's a plethora of choice from hundreds of providers covering different aspect of data ecosystem from ingestion, to logging, to stream processing, and to visualization [**?**]. Companies are tending more and more towards Cloud-native architectures for cost reduction, improved efficiency and new roles have been introduced such as chief analytics officer (CAOs) amd chief data officers (CDOs) to channel the organizational big data capabilities toward business value and competitive advantage

[**?**].

So how can one embark on this rather sophisticated journey? what can be a good logical approach to absorb the ever-increasing complexity of big data systems? how can organizations build different stacks to handle data for various workloads such as machine learning (ML), business analytics, data engineering, and streaming?

We suggest that majority of the challenge discussed starts with data architecture [**?**]; [**?**]. The data ingestion, processing and consumption of different data workloads vary, and sometimes they don't go well together. A company that enacted a data lake and a data warehouse and tries to account for both ecosystems, can be dealing with immense complexity, which in turns impact data teams, which in turn can hinder innovation, create barriers and result in monumental lost.

Development and deployment of an efficacious big data system is only the beginning of a big data journey. As data sources increase, variety of data increases, number of data consumers increase, the data store gets confuscated, and this can introduce threats for scalability and maintainability of the system. This also implies that only a handful of hyper-specialized data engineers would understand the system internals, creating silos, and potential miscommunication.

Majority of these systems are developed on-premise as ad-hoc complicated solutions that do not adhere to the practices of software engineering and software architecture [**?**]; [**?**]. As the ecosystem grows and new technologies and data processing techniques are introduced, the software architect will have a harder time to come up with a solution that address the problem requirements.

This can potentially create grounds for an immature architecture that results in solutions that are hard to scale, hard to maintain, and raise high-entry blockades [**?**]. Since the approach of ad-hoc design to big data system development is not desirable and may leave many architects and data engineers in the dark, novel data architectures that are designed specifically for BD are required. To contribute to this goal, we explore the notion of reference architectures (RAs)

and present a distributed domain-driven software RA for big data systems.

2. Why reference architecture?

To justify why we have chosen reference architectures as the suitable artefact, first we have to clarify two assumptions;

1. having a sound software architecture is essential to the successful development and maintenance of software systems
2. there exist a sufficient body of knowledge in the field of software architecture to support the development of an effective RA

One of the focal tenets of software architecture is that every system is developed to satisfy a business objective, and that the architecture of the system is a bridge between abstract business goals to concrete final solutions [**?**]. While the journey of big data can be quite challenging, the good news is that a software RA can be designed, analyzed and documented incorporating best practices, known techniques, and patterns that will support the achievement of the business goals. In this way, the complexity can be absorbed, and made tractable.

Practitioners of complex systems, software engineers, and system designers have been frequently using reference architectures to have a collective understanding of system components, functionalities, data-flows and patterns which shape the overall qualities of system and help further adjust it to the business objectives [**?**]; [**?**]. There is a fair amount of literature on reference architectures, and whereas different authors definition may vary, they all share the same tenets.

A reference architecture is amalgamation of architectural patterns, standards, software engineering techniques that bridge the problem domain to a class of solutions. This artefact can be partially or completely instantiated and prototyped in a particular business context together with other supporting artefact to enable its use. RAs are often created from previous RAs and architecture [**?**].

The usage of RAs for the development of complex systems is not new. In software product line (SPL) development, RAs are generic artifacts that are configured and instantiated for a particular domain of systems [**?**]. In software engineering, major IT giants like IBM has referred to RAs as the 'best of best practices' to address unique and complex system development challenges [**?**].

Based on the premises discussed and taking all into consideration, RAs can facilitate the issues of big data architecture and data engineering because of the following reasons;

1. RAs can promote adherence to best practice, standards, specifications and patterns
2. RAs can endow the data architecture team with openness and increase operability, incorporating architectural patterns that ensue desirable predefined quality attributes
3. RAs can be the best initial start to the big data journey, capturing design issues when they are still cheap
4. RAs can bring different stakeholders on the same table and help achieve consensus around major technological constructs
5. RAs can be effective in identifying and addressing cross-cutting concerns
6. RAs can serve as the organizational memory around design decisions, enlightening next subsequent decisions
7. RAs can act as a summary and blueprint in the portfolio of software engineers and architect, resulting in better dissemination of knowledge

3. Research Methodology

There are a few studies that have addressed the systematic development of reference architectures. Cloutier et al [**?**] present a high-level model for RA development through collection of contemporary information and capturing the essence of architectural advancements. In another effort,PuLSE-DSSA is proposed by Bayer et al. [**?**] in the context of product line development and domain engineering. PulSE-DSSA emphasizes on capturing the existing

architectural knowledge. Stricker et al. [**?**] propose a pattern-based approach for creating an RA. This study revolves around software engineering patterns motivated by the work of Gamma et al [**?**]; proposing a structural approach that includes three layers of patterns with well-defined hierarchical relationships. Nakagawa, Martins, Felizardo, and Maldonado [**?**] propose an approach to RA design outside of product line management context that is concentrated towards aspect-oriented systems.

Galster and Avgeriou [**?**] propose an empirically grounded reference architecture based on two main facets; Existing RAs in practice and available literature on RAs. Along the same vein, Nakagawa et al [**?**] presented ProSA-RA which is a 4 phase methodology that unlike many other methodologies do provide a more comprehensive instructions on RA evaluation. In addition, this methodology benefits from an ecosystem of complementary constructs that aid in RA design and evaluation such as RAModel [**?**] and a framework for evaluation of RAs (FERA) [**?**].In a recent study, Derras et al. [**?**] propose a schema of practical RA development in the context of software product line and domain engineering. This study is based on capturing knowledge from architectures in practice with attention to variability, configurability and product line development. The findings provide a four-phase process to develop quality driven reference architectures. This approach is influenced by ISO/IEC 26550 [**?**].

By analysis and study of all these approaches for design and development of RAs, a common pattern has been witnessed. Whereas some of them are more recent and some belong to years ago, there are commonalities that has been observed. All these approaches are grounded on three main pillars, 1) Existing RAs 2) RAs in literature 3) Architectures in practice. Taking this into consideration and by analyzing the results of the systematic literature review conducted by Ataei et al [**?**] we found'Empirically-grounded reference architectures' proposed by Galster and Avgeriou [**?**], a suitable methodology, because firstly it's been adopted by many studies, and secondly it's comparatively in-line with the nature of our study.

Nevertheless, we did not fully adopt this methodology and rather customized to the needs of this particular research. This is due to some inherent limitations that has been witnessed with the methodology. For instance we could not find a comprehensive guideline on how to identify data sources and how it could be categorized and synthesized into the creation of the RA in the third step of the methodology, therefore we employed the Nakagawa's information source investigation guidelines and the overall idea of the RAModel. Another limitation we've faced was with evaluation of the RA. As evaluation, second to a sound research methodology is one of the key elements of any good design science research, we had to look for a stronger and more systematic evaluation approach than what was discussed in 'empirically grounded RAs' methodology. For this purpose, and inspired by the works of Angelov et al [**?**]; [**?**], we first created an prototype of the RA in practice, and then used 'The architecture tradeoff analysis method' (ATAM) [**?**] to evaluate the artefact.

This research methodology is constituent of 6 phases which are respectively; 1) Decision on the type of the RA 2) Design strategy 3) Empirical acquisition of data 4) Construction of the RA 5) Enable RA with variability 6) Evaluation of the RA. The phrase 'empirically grounded' refers to two major elements; firstly the reference architecture should be grounded in well-established and proven principles; secondly, the reference architecture should be evaluated for applicability and validity. These don't only belong to Galster and Avgeriou methodology, and other researchers such as Cloutier [**?**] and Derras et al [**?**] have promoted the same ideas.

It is worth mentioning that this methodology is iterative, meaning that the results gained from the evaluation phase (6th phase) determines the subsequent iterations until the design reaches saturation.

3.1. Step1: Decision on type of the RA

Precursor to any effective RA development, is the decision on type of it. The type of the RA is significant, as it illuminates on information to be collected and the construction of the RA in later phases. The selection on the type of

RA for the purposes of this study is based on two dimensions; the classification framework proposed by Angelov et al. [**?**] and the usage context [**?**].

Based on the classification framework proposed by Angelov et al. [**?**], five types of RA are defined. This framework has been developed with the goal of supporting analysis of RAs with regards to context, goal, and the architecture specification/design relationships. It is based on 3 major dimensions namely context, goals, and design, each having their own corresponding sub-dimensions. These dimensions and sub-dimensions are derived by the means of interrogatives (the usage of interrogates is a well-established practice for problem analysis (the usage of interrogates is a well-established practice for problem analysis).

The interrogatives 'When', 'Where', and 'Who' have been used to address the 'context', 'Why' has been used to address 'goal', and 'How' and 'What' have been used to address 'design' dimension. The outcome of the study categorizes RAs in two major groups; 1) standardization RAs and 2) Facilitation RAs. This framework has been chosen because it is completely in-line with the purposes of this study and aims to demarcate a clear domain for the RA to be developed. The comprehensive classification of the RAs with examples in practice illuminates on how different RAs are playing roles in the industry and how they are classified. This brings clarity on what should be developed and what boundaries should be drawn.

By reading the results of the recent SLR conducted by Ataei et al on BD RAs [**?**], we've added more examples of the RAs on top of what was provided by Angelov [**?**], and provided the following updated list of RA classifications with examples;

1. Standardization RAs
   (a) Type 1: classical, standardization architectures designed to be implemented in multiple organizations. Examples are:
      i. WRM [**?**]
      ii. OSI RM [**?**]
      iii. OATH [**?**]

iv. COBRA [**?**]

v. Neomycelia [**?**]

vi. Kappa [**?**]

vii. Bolster [**?**]

(b) Type 2: classical, standardization architectures designed to be implemented in a single organization

i. Fortis Bank Reference Software Architecture [**?**]

2. Facilitation RAs

(a) Type 3: classical, facilitation reference architectures for multiple organizations designed by a software organization in cooperation with user organizations

i. Microsoft Application Architecture for .Net [**?**]

ii. IBM PanDOORA

iii. OATH [**?**]

iv. COBRA [**?**]

(b) Type 4: classical, facilitation architectures designed to be implemented in a single organization

i. Achmea Software Reference Architecture [**?**]

ii. ABN-AMRO Web Application Architecture [**?**]

(c) Type 5: preliminary, facilitation architectures designed to be implemented in multiple organizations

i. ERA [**?**]

ii. AHA [**?**]

iii. eSRA [**?**]

The domain driven distributed BD RA chosen for the purposes of this study pursues two major goals; 1) enabling and support the development and data engineering of big data systems 2) concurrently ensuring that interoperability between different heterogeneous components of the big data system is established. Therefore, the outcome artefact will be a BD RA that is a classical standardization RA designed to be implemented in multiple organizations.

3.2. Step2: Selection of Design Strategy

Angelov et al [**?**] and Galster et al[**?**] have both presented that RAs can have two major design strategies to them; 1) RAs that are designed from scratch (practice driven), 2) RAs that are based on other RAs (research driven). Designing RAs from scratch is rare, and usually takes place in an emergent domain that have not perceived a lot of attention. On the other hand, most RAs today are the amalgamation of a priori concrete architectures, models, patterns, best practices, and RAs, that together provide a compelling artefact for a class of problems.

RAs developed from scratch tend to create more prescriptive theories, whereas RAs developed based on available body of knowledge tends to provide with more descriptive design theories. The RA designed for the purposes of this study is a research-based RA based on existing RAs, concrete architectures, and best practices.

3.3. Step 3: Empirical Acquisition of Data

As aforementioned, due to the limitation witnessed by this research methodology, we have augmented this phase, and increase the systematicity and transparency of data collection and synthesis through various academic methods such as systematic literature review or SLR.

This phase is made up of three major undertakings; 1) identification of data sources; 2) capturing data sources; 3) synthesis of data sources.

3.3.1. Identification of data sources

To identify suitable data sources, we've employed the first step of ProSA-RA methodology, 'information source investigation'. This step is an endeavour to capture focal and ancillary knowledge and theories that revolve around the target domain, and lay the ground of RA development.

To unearth the architectural quanta, and to highlight gradations between various approaches to BD system development, we've selected most relevant sources as the followings;

1. **Practice-led conferences:** given that majority of recent advancements for emerging technologies such as microservices architecture [**?**];[**?**] [**?**] and big data are coming from virtually hosted practice-led conferences, we've chosen some of the best conferences hold world-wide for the purposes of data collection. These conferences are 1) Qcon [**?**] 2) State of Data Mesh by ThoughtWorks [**?**] 3) Worldwide Software Architecture Summit'21 [**?**] and 4) Kafka Summit Europe 2021 [**?**]. Our objective was to capture the frontiers of software architecture and emerging approaches currently being practiced in IT giants such as Google, Facebook and Netflix. Among all the speech in these conferences, we looked for topics that entailed the keywords 'emergent software architecture trends', 'distributed software architecture', and 'big data software architecture'. We used the software Nvivo to code the transcripts from the conference videos. We used the aforementioned keywords as the codes and associated different texts, summative, essence-capturing sentences, evocative attributes to them. During this process, a new theme 'domain driven design' emerged. We added that into the list of codes as well.
2. **Publications:** in order to capture evidence from the body of knowledge, we conducted a systematic literature review (SLR), following the guidelines of PRISMA presented by Moher et all [**?**]. The main objective of this SLR was to highlight common architectural constructs found among all the BD RAs. This SLR is build on top of our recent work [**?**] that covered all the RAs by 2020.

   The initial SLR included IEEE Explore, ScienceDirect, SpringerLink, ACM library, MIS Quarterly, Elsevier, AISel as well as citation databases such as Scopus, Web of Science, Google Scholar, and Research Gate. The SLR search keywords used were 'Big Data Reference Architectures', 'Reference Architectures in the domain of Big Data', and 'Reference Architectures and Big Data'. We followed the exact methodology, but this time for the years 2021 and 2022. Our aim was to find out if there has been any new BD RA published during the years mentioned.

By the result of this SLR, we've found 3 more BD RAs ([**?**]; [**?**]; [**?**]) and we've added two new standards ([**?**]; [**?**]) to further solidify our study. Converging these new SLR with the old, covering the years 2010-2022, we've pooled 89 literature in the primary phase, and another 10 by snowballing and citation searching. These 99 literature then went through our inclusion, exclusion and quality criteria. These criteria is as blow;

- Inclusion criteria:
  (a) studies that entailed real-world scenarios or tend to solve a problem in practice
  (b) qualitative or quantitative researches that intended to solve the industry gaps in big data system development and architecture
  (c) studies that had strong evaluations, preferably those that created the artefact in an actual organizational setup
  (d) explores the concept of RAs
  (e) provides or builds up on thorough discussion on BD RAs, limitations, drivers, and the overall ecosystem
  (f) is recent, within the years specified
  (g) is a conference paper, journal paper, book, book chapter, white paper, dissertation or thesis
- Exclusion criteria:
  (a) the study is not well evaluated or practice driven
  (b) is a duplicate
  (c) the study is not in-line with research objectives
  (d) not written in English
  (e) provides a poor quality or misleading technical information
- EQuality assessment:
  (a) is the study rich in terms of relevance to practice?
  (b) does the study create related design/design science or kernel theories?

(c) does the study entail sufficient data?

(d) does the study discuss the recent trends in BD domain?

(e) is the study based on primary data and is internationally focused?

After pooling the studies, we removed 11 studies before screening because either they were duplicates or they were not in English. The remaining 78 studies went through screen, in which, 2 studies excluded based on the exclusion criteria. From there on, 76 studies have been assessed for eligibility based on the quality framework and the inclusion criteria. The result of this process handed over 67 studies from this branch. From the other branch, 10 records identified through citation searching. These reports have been assessed through the same quality framework, inclusion and exclusion criteria, which yielded 5 studies from this stream. Together 68 studies pooled for this SLR as depicted in figure **??**.

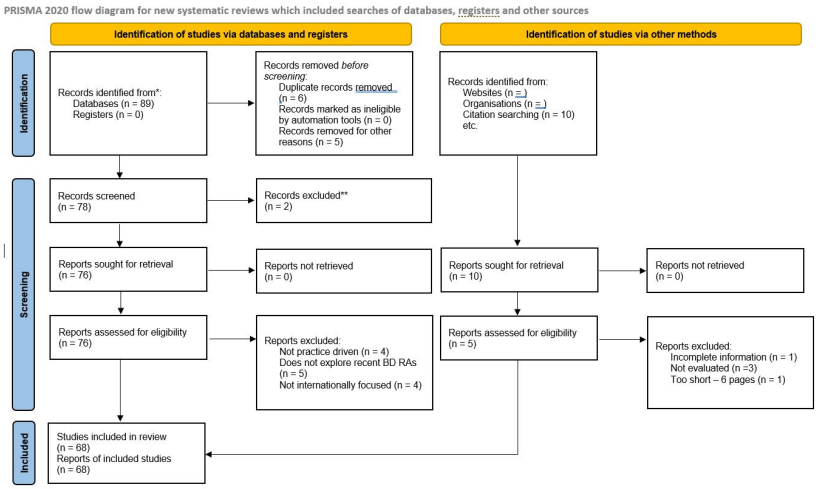

Figure 1: PRISMA flowchart

These 68 studies are comprising of journal papers, conference papers, book chapters, tech reports, tech surveys white papers, standards, master thesis, and PhD dissertations. Out of the pool of these studies, 39.4% are from IEEE Explore, 4.4% are from ScienceDirect, 23.5% are from Springerlink, 13.2% are from

ACM, and 29.4% are from other sources such as citation search, Google Scholar and Research Gate. 30 journal articles, 14 conference papers, 6 whitepapers, 2 ISO standards, 14 book chapters, and 2 postgraduate studies have been selected. 26% of these studies are from the year 2010-2013, 33% are from the years 2013-2015, and 51% are from the years 2016-2022. These stats are portrayed in figure **??**.

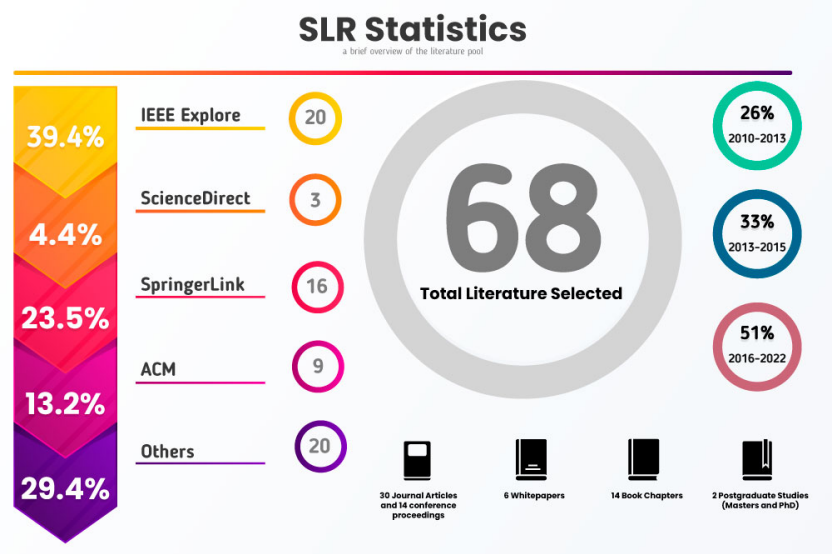

Figure 2: SLR statistics

By this stage, the research objective is set, studies are pooled, assessed and refined, thus the research embarked on the actual synthesis of data. For this purposes, the software Nvivo [**?**] has been used to code, label, and classify studies. Initially, all the keywords aforementioned has been created as nodes in the software, which are then associated to relevant sentences in studies. After coding all the studies, the findings have been synthesized to create theories, which in turn emerged themes and patterns. The findings gained from this SLR grounded the foundation for various aspect of the SLR development.

3.4. Construction of the RA

Based on the themes, theories, and patterns realized in the previous steps, the process of RA construction took place. Integral to this step was the identification of elements that the RA should contain, how these elements should be synthesized, and how the RA can be portrayed and communicated. To describe our RA, we followed ISO/IEC/IEEE 42010 standard [**?**]. This standard pivots on concrete architectures, so we did not 100% conform to it, but rather the good and relevant parts of it has been taken. For instance, architecture viewpoints, statement of corresponding rules, and expression of the architecture through architecture description languages (ADLs) have had direct aspects on the construction of this RA.

A key challenge in the development of this RA was to strike a balance between the specificity of the micro patterns and approaches to system development and general architectural concepts that reflect a view of a the system as an array of interrelated entities. Angelove et al [**?**] approached this problem by the means of interrogative through a defined framework that aims to guide the creation of RAs. Cloutier et al [**?**] suggest that a RA should entail technical, business and customer context views, whereas Vogel et al [**?**] provided classifies RA views based on the usage context, as industry specific, platform specific, industry crosscutting and product line RAs.

Stricker et al [**?**] expressed their pattern-based RA by adhering several distinct views into one. Chang et al [**?**] presented NIST BD RA as system constituent of logical components connected though interoperability interfaces in several fabrics. On the other hand, ISO/IEC/IEEE 42010 refrains from using phrases such as “technical architecture”, “physical architecture”, or “business architecture”.

Taking the best evidence from the available body of knowledge, We decided to adhere several views into one and it express the RA through a multi-layer modeling language called Archimate. Archimate is mature modeling language developed by the Open Group that provides with a uniform representation of high-level architectural diagram aimed at portraying and delineating Enterprise

architecture [**?**]. Archimate being listed as a standard architecture description language in ISO/IEC/IEEE 42010, is designed based on a set of related concepts that are specialized towards the system at different architectural layers. This means that the architect is enhanced with an integrated architectural approach that visualizes and describes different architecture domains and their underlying relations [**?**]; [**?**].

Archimate utilizes service-orientation to distinguish and relate the application, business and technology layer and use realization relationships to create relationship between concrete elements and more abstract elements across three layers. In addition, Archimate can be customized to account for varying needs of the architect.

3.5. Enabling RA with variability

Enabling RA with variability is an important process that helps with the instantiation of it. This allows RA to remain useful as a priori artefact when it comes to country-specific regulations, and organizational compliance that constrain the architect design decisions [**?**].

Variability management has been studied in the domain of Business Process Management (BPM) [**?**]; ;[**?**]; [**?**] and Software Product Line Engineering (SPLE) [**?**][**?**]; [**?**]; [**?**]; [**?**]. In BPM, variability management revolves around efficient handling of of different variants in business processes, whereas in SPLE, variability management is about modifying and extending the software artefact to account of the requirements of a specific context.

Clear identification of variability and explicit communication of it improves communication between stakeholders, allows for traceability between variation causes and effects and facilitates the decision making [**?**].

Variation points are decided based on the data collected in previous steps. Galster et al [**?**] sugguest that there are three approaches to enabling variability;

1. Annotation of the RA
2. Variability views

Whereas these approaches are enumerated, we could not find an in-detail explanation of how one should choose the appropriate variability enabling approach. Therefore, inspired by the works of Rurua et al [**?**], we decided to extend the RA with variability, by the means of Archimate annotations. The aim of this process is not find all variability points that may emerge in the usage context, but to provide with high-level system related architectural variabilities that an architect may consider for improvement of design and adoption of the RA.